ACCURACY OF RESPONSE SURFACES OVER ACTIVE SUBSPACES COMPUTED WITH RANDOM SAMPLING*

JOHN T. HOLODNAK†, ILSE C. F. IPSEN‡, AND RALPH C. SMITH§

Abstract. Given a function f that depends on m parameters, the problem is to identify an "active subspace" of dimension $k \ll m$, where f is most sensitive to change, and then to approximate f by a response surface over this lower-dimensional active subspace.

We present a randomized algorithm for determining k and computing an orthonormal basis for the active subspace. We also derive a tighter probabilistic bound on the number of samples required for approximating the active subspace to a user-specified accuracy. The bound does not explicitly depend on the total number m of parameters, and allows tuning of the failure probability. We discuss different error measures for response surfaces; and separate errors due to approximation over a subspace from errors due to construction of the response surface. The accuracy of the construction method for the response surface is of utmost importance. We design a test problem that makes it easy to construct active subspaces of any dimension k. Numerical experiments with $k = 10$ and a response surface constructed with sparse grid interpolation confirm the effectiveness of our error measures.

Key words. interpolation, sparse grids, singular value decomposition, randomized algorithms, concentration inequalities

AMS subject classification. 15A18, 15A23, 15A60, 15B10, 35J25, 60G60, 65N30, 65C06, 65C30, 65F15, 65D05

1. Introduction.

A differentiable function $f : \mathbb{R}^m \to \mathbb{R}$ that is expensive to evaluate can be approximated by a *response surface* h, which is a function that is "close" to f in some sense but much cheaper to evaluate. Such response surfaces can be constructed by evaluating f at a number of *training points*, and then fitting a surface at the training points. If m is large, however, a good approximation for f may require many training points.

One can make the construction of the response surface cheaper by dimension reduction, that is, by identifying a lower-dimensional *active subspace* in the parameter space $\mathbb{R}^m$ [6, 24]. The active subspace represents linear combinations of the m parameters along which f is most sensitive. An orthogonal basis for an active subspace[1] can be computed from k dominant eigenvectors of a Monte Carlo approximation to the $m \times m$ matrix

$$E = \int_{\mathbb{R}^m} \nabla f(\mathbf{x}) \left(\nabla f(\mathbf{x})\right)^T \rho(\mathbf{x})\, d\mathbf{x}.$$

*The second author was supported in part by NSF grant CCF-1145383. The second author also acknowledges the support from the XDATA Program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323 FA8750-12-C-0323. The third author was supported in part by the Consortium for Advanced Simulation of Light Water Reactors (`http://www.casl.gov`), an Energy Innovation Hub (`http://www.energy.gov/hubs`) for Modeling and Simulation of Nuclear Reactors under U.S. Department of Energy Contract No. DE-AC05-00OR22725.

†Work conducted while a student at North Carolina State University, (`jtholodn@ncsu.edu`)

‡Department of Mathematics, North Carolina State University, P.O. Box 8205, Raleigh, NC 27695-8205, USA (`ipsen@ncsu.edu`, `http://www4.ncsu.edu/~ipsen/`)

§Department of Mathematics, North Carolina State University, P.O. Box 8205, Raleigh, NC 27695-8205, USA (`rsmith@ncsu.edu`, `http://www4.ncsu.edu/~rsmith/`)

[1]Since active subspaces are highly non-unique, we use the indefinite article, unless we have in mind a specific subspace.

If $k \ll m$, then constructing a response surface from this k-dimensional active subspace is much cheaper than constructing it from the full parameter space $\mathbb{R}^m$.

The concept of "active subspace" was introduced by Russi [24, Chapter 6], and formalized by Constantine et al. [6]. Since they rely on the eigendecomposition of a covariance matrix, active subspaces are related to principal component analysis [17].

Active subspaces have been applied to the solution of mathematical problems, including stochastic PDEs [6, 9, 30], and reduced-order nonlinear models [2].

Active subspaces have also found applications in engineering. In [4], two functions related to the manufacturing error of airfoils, both of which depend on twenty variables, are approximated by a response surface over a one-dimensional active subspace. In [5], a function related to the wall pressure of combustors, that depends on six variables, is approximated over an active subspace of three variables. In [3], a model of an annular combustor in 38 variables is approximated with only three variables. In [10], an active subspace of one dimension is identified for the power of a photovoltaic cell, which depends on five variables. In [21], active subspaces are combined with kriging to construct response surfaces for airfoil design problems.

1.1. Our contributions. The paper contains three major contributions:

1. A tighter bound (Theorem 3.1) for the number of Monte Carlo samples required to approximate an active subspace to a user-specified error.
   The bound does not explicitly depend on the total number m of parameters.
2. Different error measures (Section 4) to quantify the accuracy of a response surface h.
   We carefully distinguish the error due to the construction of h from the error due to approximating f over an active subspace. We emphasize that if care is not taken to construct a good response surface h over the active subspace, then h does a poor job of approximating f everywhere.
3. Design of a simple test problem (Section 5.3) to produce active subspaces of any dimension k.
   Numerical experiments (Section 6) on the test problem demonstrate that a function of 3495 variables can be approximated over an active subspace of dimension $k = 10$ with relative accuracy.

1.2. Outline. Section 2 presents a review of ideal active subspaces and response surfaces, algorithms for their approximation, along with error bounds. In Section 3, we derive a tighter bound on the number of samples required to approximate the active subspace to within a user-specified error, with the proofs relegated to Appendix A. Section 4 contains a discussion of several error measures for approximate response surfaces. In Section 5, we extend a test problem with a one-dimensional active subspace to a class of problems with active subspaces of any dimension. The numerical experiments in Section 6 apply the error measures from Section 4 to a test problem from Section 5 with a 10-dimensional active subspace. Section 7 presents a brief summary, and makes a recommendation for how to compute high-dimensional active subspaces.

1.3. Assumptions and Notation. Column vectors with m real elements are represented by lower case bold face letters, such as $\mathbf{x} \in \mathbb{R}^m$. The Euclidean norm is $\|\mathbf{x}\|_2 \equiv \sqrt{\mathbf{x}^T\mathbf{x}}$, where the superscript T denotes the transpose. Upper case roman letters, such as E, denote real matrices; and I is an identity matrix whose dimension is clear from the context.

The set $\mathcal{N}(0, 1)$ represents normally distributed random variables with mean 0

and variance 1, while $\mathcal{U}(a,b)$ represents uniformly distributed random variables in the interval $[a,b]$.

Upper case bold face letters, such as $\mathbf{X}$, represent random vectors with probability density functions $\rho(\mathbf{x})$. The expected value of a function $h(\mathbf{X}) : \mathbb{R}^m \to \mathbb{R}$ with respect to $\mathbf{X}$ is

$$\mathbf{E}_{\mathbf{X}}\left[h(\mathbf{X})\right] \equiv \int_{\mathbb{R}^m} h(\mathbf{x})\,\rho(\mathbf{x})\,d\mathbf{x}.$$

The following assumptions hold throughout the remainder of the paper.

ASSUMPTIONS 1.1. *The vector $\mathbf{X} \in \mathbb{R}^m$ is a random vector with probability density function $\rho(\mathbf{x})$, which is positive, $\rho(\mathbf{x}) > 0$ and bounded for all $\mathbf{x} \in \mathbb{R}^m$.*

The function $f(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable. Hence f is Lipschitz continuous, and there exists $L > 0$ with $\|\nabla f(\mathbf{x})\|_2 \leq L$ for all $\mathbf{x} \in \mathbb{R}^m$.

The positivity of $\rho(\mathbf{x})$ ensures that the conditional probability density functions in Sections 2.1 and 2.2 are well-defined, while the boundedness ensures that we can integrate with respect to $\rho(\mathbf{x})$.

2. Review of active subspace identification and response surface construction. We review results from [6, 7, 8] about identification of active subspaces and construction of response surfaces.

Section 2.1 introduces ideal active subspaces and response surfaces, while Section 2.2 presents sampling-based algorithms for their approximation. Then we review existing error bounds for: the mean squared error of the approximate response surface in Section 2.3, and the asymptotic number of samples required to compute an active subspace to a user-specified accuracy in Section 2.4.

2.1. Ideal active subspaces and response surfaces. We review the definition of active subspaces from [6] and how to determine ideal active subspaces and response surfaces.

Ideal active subspaces. The sensitivity of f along a direction $\mathbf{v}$, with $\|\mathbf{v}\|_2 = 1$, can be estimated from the expected value of the squared directional derivative of f along $\mathbf{v}$,

$$\mathbf{E}_{\mathbf{X}}\left[\left(\mathbf{v}^T\,\nabla f(\mathbf{X})\right)^2\right] = \int_{\mathbb{R}^m} \left(\mathbf{v}^T\,\nabla f(\mathbf{x})\right)^2 \rho(\mathbf{x})\,d\mathbf{x}.$$

Mean squared derivatives also appear in [26] for measuring sensitivity, but along coordinate directions. Directional derivatives, in contrast, are not just confined to coordinate directions but can measure sensitivity in any direction.

The following result associates expected values of squared directional derivatives along certain directions with eigenvalues and eigenvectors of a matrix.

LEMMA 2.1 (Lemma 2.1 in [6]). *Let Assumptions 1.1 hold, and define the $m \times m$ matrix*

$$E \equiv \int_{\mathbb{R}^m} \nabla f(\mathbf{x})\left(\nabla f(\mathbf{x})\right)^T \rho(\mathbf{x})\,d\mathbf{x}. \tag{2.1}$$

If $E\mathbf{v} = \lambda\,\mathbf{v}$ for a scalar λ and a vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 = 1$, then

$$\mathbf{E}_{\mathbf{X}}\left[(\mathbf{v}^T\,\nabla f(\mathbf{X}))^2\right] = \lambda.$$

Proof. This follows from $\mathbf{E}_{\mathbf{X}}\left[(\mathbf{v}^T \nabla f(\mathbf{X}))^2\right] = \mathbf{v}^T E \mathbf{v} = \lambda\, \mathbf{v}^T \mathbf{v} = \lambda$. ∎

Since the $m \times m$ matrix E is symmetric positive semi-definite, it has an eigendecomposition

$$E = V \Lambda V^T, \qquad \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix} \quad \text{with} \quad \lambda_1 \geq \cdots \geq \lambda_m \geq 0, \tag{2.2}$$

and $V \in \mathbb{R}^{m \times m}$ is an orthogonal matrix of eigenvectors.

Lemma 2.1 implies that eigenvectors associated with dominant eigenvalues represent directions along which f is most sensitive. Dominant eigenvalues can be identified from large relative eigenvalue gaps [8, Section 4.1].

DEFINITION 2.2 (Ideal active subspace). *If E has k dominant eigenvalues with $\lambda_k \gg \lambda_{k+1}$, partition the eigendecomposition (2.2) conformally*

$$\Lambda = \begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix}, \quad \textit{with} \quad \Lambda_1 = \operatorname{diag}\begin{pmatrix} \lambda_1 & \cdots & \lambda_k \end{pmatrix}, \qquad V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}. \tag{2.3}$$

The orthonormal columns of the $m \times k$ matrix V_1 are the eigenvectors associated with the k dominant eigenvalues of E.

We call $\mathsf{range}(V_1)$ *the* ideal active subspace *of dimension k for f.*

Ideal response surfaces. Now we construct a response surface over the ideal active subspace $\mathsf{range}(V_1)$. To this end set

$$\mathbf{X} = V V^T \mathbf{X} = V_1 \mathbf{Y} + V_2 \mathbf{Z}, \qquad \text{where} \qquad \mathbf{Y} \equiv V_1^T \mathbf{X}, \quad \mathbf{Z} \equiv V_2^T \mathbf{X}.$$

Here $\mathbf{Y} \in \mathbb{R}^k$ represents the coordinates of the projection of $\mathbf{X}$ onto the active subspace $\mathsf{range}(V_1)$, while $\mathbf{Z} \in \mathbb{R}^{m-k}$ represents the coordinates of the projection of $\mathbf{X}$ onto the orthogonal complement. The purpose of this decomposition is to approximate $f(\mathbf{X})$ by a function that depends on $\mathbf{Y}$ only.

The simplest approximation would be[2] $f(V_1 \mathbf{Y}) = f(V_1 V_1^T \mathbf{X})$. In the special case $\lambda_{k+1} = \cdots = \lambda_m = 0$, the function f is constant over $\mathsf{range}(V_2)$ and the approximation is exact, $f(V_1 \mathbf{y}) = f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^m$.

In general, though, $\lambda_{k+1} > 0$ and f changes over $\mathsf{range}(V_2)$. We can improve the approximation $f(V_1 \mathbf{Y})$ by averaging over all $\mathbf{z} \in \mathbb{R}^{m-k}$.

DEFINITION 2.3 (Ideal response surface). *We define the* ideal response surface *of $f(\mathbf{X}) = f(V_1 \mathbf{Y} + V_2 \mathbf{Z})$ associated with the ideal active subspace* $\mathsf{range}(V_1)$ *as the conditional expectation of $f(\mathbf{X})$ given $\mathbf{Y} = \mathbf{y}$,*

$$g(\mathbf{Y}) \equiv \mathbf{E}_{\mathbf{Z}}\left[f(V_1 \mathbf{Y} + V_2 \mathbf{Z}) \mid \mathbf{Y} = \mathbf{y}\right] = \int_{\mathbb{R}^{m-k}} f(V_1 \mathbf{y} + V_2 \mathbf{z})\, \rho_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z} \,|\, \mathbf{y})\, d\mathbf{z}. \tag{2.4}$$

Here $g : \mathbb{R}^k \to \mathbb{R}$; and $\rho_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z} \,|\, \mathbf{y})$ is the conditional density of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$.

Note that $\mathbf{X} = V_1 \mathbf{Y} + V_2 \mathbf{Z}$. So the joint density of $\mathbf{Y}$ and $\mathbf{Z}$ is $\rho(\mathbf{y}, \mathbf{z}) = \rho(\mathbf{x})$; and the marginal density of $\mathbf{Y}$ is $\rho_{\mathbf{Y}}(\mathbf{y}) \equiv \int_{\mathbb{R}^{m-k}} \rho(\mathbf{y}, \mathbf{z})\, d\mathbf{z}$. Thus [23, Section 3.3] the density in (2.4) can be written as

$$\rho_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z} \,|\, \mathbf{y}) = \rho(\mathbf{y}, \mathbf{z}) / \rho_{\mathbf{Y}}(\mathbf{y}).$$

[2] Since f is defined on all of $\mathbb{R}^m$, the projection $V_1 V_1^T \mathbf{x}$ is automatically in the domain. However, if the domain of f is a proper subset of $\mathbb{R}^m$, then $V_1 V_1^T \mathbf{x}$ may not be in the domain. See [30, Section 1.1] and [6, Section 4.1.2] for a discussion of this issue.

Division by zero does not occur because Assumptions 1.1 guarantee $\rho(\mathbf{x}) > 0$ on $\mathbb{R}^m$.

REMARK 2.4. *The ideal response surface (2.4) is optimal in the sense of the mean squared error [13, Section 7.9: Theorem 17].*

Specifically, among all functions $\phi(\mathbf{Y}) : \mathbb{R}^k \to \mathbb{R}$ *with* $\mathbf{E}_\mathbf{Y}\left[\phi(\mathbf{Y})^2\right] < \infty$, *the function that minimizes*

$$\mathbf{E}_\mathbf{Z}\left[\left(f(V_1\mathbf{Y} + V_2\mathbf{Z}) - \phi(\mathbf{Y})\right)^2 \mid \mathbf{Y} = \mathbf{y}\right]$$

is

$$g(\mathbf{Y}) = \mathbf{E}_\mathbf{Z}\left[f(V_1\mathbf{Y} + V_2\mathbf{Z}) \mid \mathbf{Y} = \mathbf{y}\right] \;=\; \mathbf{E}_\mathbf{Z}\left[f(\mathbf{X}) \mid \mathbf{Y} = \mathbf{y}\right].$$

The mean squared error over all $\mathbf{x}$, which results from approximating $f(\mathbf{X})$ with the ideal response (2.4), can be bounded by the subdominant eigenvalues of E.

THEOREM 2.5 (Theorem 3.1 in [6]). *Let Assumptions 1.1 hold. If* E *in (2.1) has an eigendecomposition partitioned as in (2.3) with* $\lambda_k > \lambda_{k+1}$, *then*

$$\int_{\mathbb{R}^m} \Big(f(\mathbf{X}) - \mathbf{E}_\mathbf{Z}\left[f(\mathbf{X}) \mid \mathbf{Y} = \mathbf{y}\right]\Big)^2 \rho(\mathbf{x})\, d\mathbf{x} \le c_1(\lambda_{k+1} + \cdots + \lambda_m),$$

where c_1 *is a constant that depends on* $\rho(\mathbf{x})$.

Thus, if the subdominant eigenvalues are small, then the ideal response surface (2.4) is a good approximation to f.

2.2. Approximate active subspaces and response surfaces. Computing the matrix E in (2.1) and the ideal response surface in (2.4) is expensive because they involve integrals over high-dimensional spaces.

Approximate active subspaces. We approximate E with a Monte Carlo method as in [6, (2.16)],

$$\widehat{E} \equiv \frac{1}{n_1}\sum_{i=1}^{n_1} \nabla f(\mathbf{x}_i)\left(\nabla f(\mathbf{x}_i)\right)^T$$

where the vectors $\mathbf{x}_i \in \mathbb{R}^m$ are sampled according to $\rho(\mathbf{x})$, and $n_1 \le m$ is the number of samples, determined as in Sections 2.4 and 3. The $m \times m$ matrix $\widehat{E}$ is symmetric positive semi-definite with eigendecomposition

$$\widehat{E} = \widehat{V}\widehat{\Lambda}\widehat{V}^T, \qquad \widehat{\Lambda} = \begin{pmatrix} \widehat{\lambda}_1 & & \\ & \ddots & \\ & & \widehat{\lambda}_m \end{pmatrix} \quad \text{where} \quad \widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_m \ge 0, \qquad (2.5)$$

and $\widehat{V} \in \mathbb{R}^{m\times m}$ is an orthogonal matrix of eigenvectors.

We assume that the approximation $\widehat{E}$ has preserved the large relative eigenvalue gaps from E, and use the same k below as in Definition 2.2.

DEFINITION 2.6 (Approximate active subspace). *If* $\widehat{E}$ *has* k *dominant eigenvalues with* $\widehat{\lambda}_k > \widehat{\lambda}_{k+1}$, *partition the eigendecomposition (2.5) conformally*

$$\widehat{\Lambda} = \begin{pmatrix} \widehat{\Lambda}_1 & \\ & \widehat{\Lambda}_2 \end{pmatrix}, \quad \textit{with} \quad \widehat{\Lambda}_1 = \operatorname{diag}\begin{pmatrix} \widehat{\lambda}_1 & \cdots & \widehat{\lambda}_k \end{pmatrix}, \qquad \widehat{V} = \begin{pmatrix} \widehat{V}_1 & \widehat{V}_2 \end{pmatrix}, \; (2.6)$$

where the orthonormal columns of the $m \times k$ *matrix* $\widehat{V}_1$ *are the eigenvectors associated with the* k *dominant eigenvalues of* $\widehat{E}$.

We call $\mathsf{range}(\widehat{V}_1)$ *the* approximate active subspace *of dimension* k *for* f.

Algorithm 1 Computing an approximate active subspace of f

Input:

- Functions $f(\mathbf{x})$ and $\rho(\mathbf{x})$ satisfying Assumptions 1.1
- Integer $0 < n_1 \leq m$ {Number of columns in G}

Output:

- $m \times k$ matrix $\widehat{V}_1$ with orthonormal columns
  {Basis for approximate active subspace}

$G = \mathbf{0}_{m \times n_1}$
for $i = 1 : n_1$ **do**
  Sample $\mathbf{x}_i \in \mathbb{R}^m$ according to $\rho(\mathbf{x})$
  $G(:, i) = \nabla f(\mathbf{x}_i)$
end for
Compute the thin SVD $G = U \Sigma W^T$ in (2.7)
Choose an integer k with $\sigma_k^2 \gg \sigma_{k+1}^2$
$\widehat{V}_1 = U(:, 1 : k)$

Computation. One can avoid the explicit computation of $\widehat{E}$ by representing it in factored form, as a Gram matrix,

$$\widehat{E} = \frac{1}{n_1} G G^T, \qquad \text{where} \quad G \equiv \begin{pmatrix} \nabla f(\mathbf{x}_1) & \dots & \nabla f(\mathbf{x}_{n_1}) \end{pmatrix}.$$

The eigenvalues and eigenvectors of $\widehat{E}$ are then computed from a thin singular value decomposition of the $m \times n_1$ matrix G.

Let $n_1 \leq m$ and $r \equiv \mathsf{rank}\,(G) = \mathsf{rank}\left(\widehat{E}\right)$. Then G has a thin singular value decomposition (SVD)

$$G = U \Sigma W^T, \qquad \Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \quad \text{where} \quad \sigma_1 \geq \cdots \geq \sigma_r > 0, \tag{2.7}$$

and $U \in \mathbb{R}^{m \times r}$ and $W \in \mathbb{R}^{n_1 \times r}$ have orthonormal columns.

The left singular vectors U of G span the column space of $\widehat{E}$, while the singular values of G are related to the eigenvalues of $\widehat{E}$ by $\frac{1}{n_1}\sigma_i^2 = \widehat{\lambda}_i$, $1 \leq i \leq r$. Algorithm 1 summarizes the computation of the approximate active subspace.

Approximate response surfaces. We start by determining the ideal response surface for the approximate active subspace. Then we apply two conceptual approximation steps, followed by an interpolation to determine an approximate response surface.

As for the ideal response surface, start with

$$\mathbf{X} = \widehat{V}\widehat{V}^T \mathbf{X} = \widehat{V}_1 \widehat{\mathbf{Y}} + \widehat{V}_2 \widehat{\mathbf{Z}}, \qquad \text{where} \qquad \widehat{\mathbf{Y}} \equiv \widehat{V}_1^T \mathbf{X}, \quad \widehat{\mathbf{Z}} \equiv \widehat{V}_2^T \mathbf{X}.$$

Then the conditional expectation of $f(\mathbf{X}) = f(\widehat{V}_1 \widehat{\mathbf{Y}} + \widehat{V}_2 \widehat{\mathbf{Z}})$ given $\widehat{\mathbf{Y}} = \widehat{\mathbf{y}}$ is

$$\begin{aligned} \hat{g}(\widehat{\mathbf{Y}}) &\equiv \mathbf{E}_{\widehat{\mathbf{Z}}} \left[f\left(\widehat{V}_1 \widehat{\mathbf{Y}} + \widehat{V}_2 \widehat{\mathbf{Z}}\right) \mid \widehat{\mathbf{Y}} = \widehat{\mathbf{y}} \right] \\ &= \int_{\mathbb{R}^{m-k}} f\left(\widehat{V}_1 \widehat{\mathbf{y}} + \widehat{V}_2 \widehat{\mathbf{z}}\right) \rho_{\widehat{\mathbf{Z}}|\widehat{\mathbf{Y}}}\left(\widehat{\mathbf{z}} \mid \widehat{\mathbf{y}}\right) d\widehat{\mathbf{z}}. \end{aligned} \tag{2.8}$$

Algorithm 2 Computing training pairs for the approximate response surface

Input:

- Functions $f(\mathbf{x})$ and $\rho(\mathbf{x})$ satisfying Assumptions 1.1
- $m \times k$ matrix $\widehat{V}_1$ with orthonormal columns
  {Basis for approximate active subspace}
- Integer $n_2 > 0$ {Number of samples for Monte Carlo integration}
- Training points $\widehat{\mathbf{y}}_i \in \mathbb{R}^k$, $1 \le i \le T$

Output:

Training pairs $(\widehat{\mathbf{y}}_i, \widehat{\mathbf{t}}_i)$, $1 \le i \le T$

for $i = 1 : T$ **do**
  Sample $\widehat{\mathbf{z}}_j \in \mathbb{R}^{m-k}$ according to $\rho_{\widehat{\mathbf{Z}}\,|\,\widehat{\mathbf{Y}}}$, $1 \le j \le n_2$
  $\widehat{\mathbf{t}}_i = \sum_{j=1}^{n_2} f(\widehat{V}_1\widehat{\mathbf{y}}_i + \widehat{V}_2\widehat{\mathbf{z}}_j)$
end for

is the ideal response surface for the approximate subspace $\mathsf{range}(\widehat{V}_1)$.

Now we go about the approximation. First, replace the high-dimensional integrals (2.8) with Monte Carlo approximations

$$\widetilde{g}(\widehat{\mathbf{y}}) \equiv \frac{1}{n_2}\sum_{j=1}^{n_2} f\left(\widehat{V}_1\widehat{\mathbf{y}} + \widehat{V}_2\widehat{\mathbf{z}}_j\right) \tag{2.9}$$

where the vectors $\widehat{\mathbf{z}}_j \in \mathbb{R}^{m-k}$ are sampled according to the conditional density function $\rho_{\widehat{\mathbf{Z}}\,|\,\widehat{\mathbf{Y}}}$, and n_2 is the number of samples. Second, compute the Monte Carlo approximations (2.9) only at T training points $\widehat{\mathbf{y}}_i$,

$$\widehat{\mathbf{t}}_i \equiv \widetilde{g}(\widehat{\mathbf{y}}_i), \qquad 1 \le i \le T.$$

Algorithm 2 summarizes the corresponding computations for user-specified training points.

DEFINITION 2.7 (Approximate response surface). *The* approximate response surface *for the approximate active subspace* $\mathsf{range}(\widehat{V}_1)$ *is a function* $h : \mathbb{R}^k \to \mathbb{R}$ *that interpolates* $\widetilde{f}$ *at the* T *training pairs* $(\widehat{\mathbf{y}}_i, \widehat{\mathbf{t}}_i)$,

$$h(\widehat{\mathbf{y}}_i) = \widehat{\mathbf{t}}_i, \qquad 1 \le i \le T.$$

In Section 6, we construct h from piecewise multilinear interpolation over a sparse grid.

2.3. Error of the approximate response surface. We present a bound on the absolute error of the approximate response surface when the approximate active subspace is computed by Algorithm 1 and the training pairs by Algorithm 2.

No assumptions are made on the particular interpolation method for h, other than a bound on the mean squared error between $h(\widehat{\mathbf{y}})$ and $\widetilde{g}(\widehat{\mathbf{y}})$ for all $\widehat{\mathbf{y}} \in \mathbb{R}^k$.

Theorem 2.8 is a restatement of [6, Theorem 3.7] and bounds the mean squared error between h and f.

THEOREM 2.8 (Theorem 3.7 in [6], [7]). *Suppose that*

1. *$f(\mathbf{x})$ and $\rho(\mathbf{x})$ satisfy Assumptions 1.1.*
2. *The eigenvalues of E in Lemma 2.1 have a gap $\lambda_k > \lambda_{k+1}$ for some $1 \le k < m$.*
3. *The ideal active subspace V_1 is given in Definition 2.2.*
4. *Algorithm 1 computes a $m \times k$ matrix $\widehat{V}_1$ with orthonormal columns so that $\left\| V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T \right\|_2 \le \epsilon$ for some $\epsilon > 0$.*
5. *Algorithm 2 computes $\widehat{\mathbf{t}}_i = \widetilde{g}(\mathbf{y}_i)$ from n_2 samples for each $1 \le i \le T$.*
6. *The approximate response surface h interpolates the T training pairs $(\widehat{\mathbf{y}}_i, \widehat{\mathbf{t}}_i)$ such that for some $c_2 > 0$ and all $\widehat{\mathbf{y}} \in \mathbb{R}^k$*

$$\mathbf{E}_{\mathbf{Z}}\left[\left(\widetilde{g}(\widehat{\mathbf{Y}}) - h(\widehat{\mathbf{Y}})\right)^2 \mid \widehat{\mathbf{Y}} = \widehat{\mathbf{y}}\right] \le c_2.$$

Then, with the $\rho(\mathbf{x})$-dependent constant c_1 from Lemma 2.1,

$$\int_{\mathbb{R}^m} \left(f(\mathbf{x}) - h(\widehat{\mathbf{y}})\right)^2 \rho(\mathbf{x})\, d\mathbf{x} \;\le\; \left[\sqrt{c_1}\left(1 + \frac{1}{\sqrt{n_2}}\right)\left(\epsilon\sqrt{\lambda_1 + \cdots + \lambda_k} + \sqrt{\lambda_{k+1} + \cdots + \lambda_m}\right) + \sqrt{c_2}\right]^2.$$

REMARK 2.9. *We make the following observations about Theorem 2.8.*

1. *The mean squared error in item 6 is equal to the integral*

$$\mathbf{E}_{\mathbf{Z}}\left[\left(\widetilde{g}(\widehat{\mathbf{Y}}) - h(\widehat{\mathbf{Y}})\right)^2 \mid \widehat{\mathbf{Y}} = \widehat{\mathbf{y}}\right] = \int_{\mathbb{R}^{m-k}} \left(\widetilde{g}(\widehat{\mathbf{y}}) - h(\widehat{\mathbf{y}})\right)^2 \rho_{\widehat{\mathbf{Z}}|\widehat{\mathbf{Y}}}(\widehat{\mathbf{z}} \,|\, \widehat{\mathbf{y}})\, d\mathbf{z}.$$

2. *The mean squared error in the approximate response surface h is small if:*
   (a) *The eigenvalues $\lambda_{k+1}, \ldots, \lambda_m$ of E are small;*
   (b) *The approximate active subspace spanned by $\widehat{V}_1$ has the same dimension as and is close to the ideal active subspace spanned by V_1, see Section 2.4;*
   (c) *The number of Monte Carlo samples n_2 for the training points is sufficiently large;*
   (d) *The interpolation for computing the approximate response surface h is sufficiently accurate.*
3. *The accuracy of the interpolation method for h is crucial. For the error in h to be small, c_2 must be small since it appears as an additive term in the error bound.*
4. *The number of Monte Carlo samples n_2 has only a weak influence on the error, since $1 \le (1 + \frac{1}{\sqrt{n_2}}) \le 2$.*

2.4. Bounding the angle between approximate and ideal subspaces. To capture item 3 in Theorem 2.8, we review a probabilistic bound from [8] on the asymptotic number of samples n_1 required to compute an approximate active subspace to a user-specified accuracy.

Again, we assume that the approximate active subspace has the same dimension as the ideal active subspace. Since the $m \times k$ matrices V_1 and $\widehat{V}_1$ have orthonormal columns, the orthogonal projectors onto the ideal and approximate active subspaces are $V_1 V_1^T$ and $\widehat{V}_1 \widehat{V}_1^T$, respectively. Let θ be the largest principal angle between the approximate and ideal active subspaces then [12, Sections 2.5.3, 6.4.3] and [28, Corollary

2.6]

$$\sin\theta = \left\| V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T \right\|_2 . \tag{2.10}$$

Thus, the sine of the largest principal angle equals the two-norm difference between the projectors. The probabilistic bound below specifies the asymptotic number of samples required to limit the angle to a user-specified accuracy.

THEOREM 2.10 (Corollary 3.7 in [8]). *Suppose that*

1. *$f(\mathbf{x})$ and $\rho(\mathbf{x})$ satisfy Assumptions 1.1.*
2. *The eigenvalues of the $m \times m$ matrix E in Lemma 2.1 have a gap $\lambda_k > \lambda_{k+1}$ for some $1 \leq k < m$.*
3. *The ideal active subspace of dimension k is given in Definition 2.2.*
4. *Algorithm 1 computes a $m \times k$ matrix $\widehat{V}_1$ with orthonormal columns.*
5. *The desired accuracy is limited by the relative gap $0 < \epsilon < \frac{\lambda_k - \lambda_{k+1}}{5\,\lambda_1}$.*

If the number of samples in Algorithm 1 is

$$n_1 = \Omega\left(\max\left\{ \frac{L^2}{\lambda_1\,\epsilon}, \frac{\nu^2}{\lambda_1^2\,\epsilon^2} \right\} \ln(2\,m) \right),$$

where ν^2 is a measure of variance [8, (3.25)], then with high probability

$$\sin\theta = \left\| V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T \right\|_2 \leq \frac{4\,\lambda_1\,\epsilon}{\lambda_k - \lambda_{k+1}}.$$

Theorem 2.10 depends on the dimension m of E, which is also the number of parameters. We remove this dependence in a tighter non-asymptotic bound in the next section.

3. A tighter bound on the angle between approximate and ideal active subspaces. We present a tighter bound on the angle between approximate and ideal subspaces that does not depend explicitly on the total number m of parameters and allows tuning of the failure probability.

THEOREM 3.1. *Suppose that*

1. *$f(\mathbf{x})$ and $\rho(\mathbf{x})$ satisfy Assumptions 1.1.*
2. *The eigenvalues of the $m \times m$ matrix E in Lemma 2.1 have a gap $\lambda_k > \lambda_{k+1}$ for some $1 \leq k < m$.*
3. *The ideal active subspace of dimension k is given in Definition 2.2.*
4. *Algorithm 1 computes a $m \times k$ matrix $\widehat{V}_1$ with orthonormal columns.*
5. *The desired accuracy is limited by the relative gap $0 < \epsilon < \frac{\lambda_k - \lambda_{k+1}}{4\,\lambda_1}$.*
6. *$0 < \delta < 1$ is a user-specified failure probability.*

If the number of samples in Algorithm 1 is at least

$$n_1 \geq \frac{8}{3}\,\frac{L^2}{\lambda_1\,\epsilon^2}\,\ln\left(\frac{4}{\delta}\,\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1} \right),$$

then with probability at least $1 - \delta$,

$$\sin\theta = \left\| V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T \right\|_2 \leq \frac{4\,\lambda_1\,\epsilon}{\lambda_k - \lambda_{k+1}}.$$

Proof. See Section A. □

Like Theorem 2.10, Theorem 3.1 is conceptual in the sense that it require knowledge of the eigenvalues of E and a global bound L on $\|\nabla f(\mathbf{x})\|_2$. Nevertheless, Theorem 3.1 is informative because it implies that approximating an active subspace is easier if:

1. The relative eigenvalue gap $(\lambda_k - \lambda_{k+1})/\lambda_1$ is large.
   One can view the inverse of this relative gap as a measure of sensitivity for the computed subspace $\widehat{V}_1$. This measure is invariant under scalar multiplication of E.
2. The function f is smooth.
3. The matrix E has small *intrinsic dimension*[3]

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1} = \frac{\mathsf{trace}\,(E)}{\|E\|_2},$$

   meaning E has low numerical rank.

4. Two error measures for the approximate response surface. We discuss two ways of measuring the error of the approximate response surface h as an approximation of f.

4.1. Mean squared error. Theorem 2.8 bounds an absolute error that is averaged over the whole $\mathbb{R}^m$. It does not distinguish the error over the active subspace, which one would expect to be considerably smaller, from the error outside the active subspace.

4.2. Pointwise relative error. A more precise measure is the relative error between h and f

$$\left| \frac{h(\widehat{\mathbf{y}}) - f(\mathbf{x})}{f(\mathbf{x})} \right|, \qquad \text{where} \quad \widehat{\mathbf{y}} = \widehat{V}_1^T \mathbf{x}. \tag{4.1}$$

for different points $\mathbf{x}$. We distinguish errors (4.1) at points $\mathbf{x}$ in the approximate active subspace $\mathsf{range}(\widehat{V}_1)$ from those outside the approximate active subspace.

If h is not sufficiently accurate in the approximate active subspace, i.e. large errors (4.1) for $\mathbf{x} \in \mathsf{range}(\widehat{V}_1)$, then h is unlikely to be accurate at other points as well. it is important to construct a response surface in such a way as to ensure some degree of accuracy.

Even if h is sufficiently accurate over $\mathsf{range}(\widehat{V}_1)$, the errors (4.1) can still be large for $\mathbf{x}$ outside the active subspace. This can happen if $\lambda_{k+1}, \ldots, \lambda_m$ are too large, or if $\widehat{V}_1$ is not a good approximation to V_1.

5. A specific problem. To demonstrate the effectiveness of our approach in Section 4.2, we generate active subspaces of higher dimensions by generalizing a problem from [6, Section 5].

We define Gaussian random fields (Section 5.1) before describing the original problem (Section 5.2) and our generalization (Section 5.3).

5.1. Gaussian random fields. Since the original problem [6, Section 5] and our generalization involve log-Gaussian random fields, we present brief definitions, but limit them strictly to our very specific context.

[3] See Definition A.1.

DEFINITION 5.1 (Definition 1.3 in [1]). *A* Gaussian random field $a(\mathbf{s})$ *over* $\mathbb{R}^2$ *is a scalar function such that, for any integer* $k > 0$*, and any* k *fixed points* $\mathbf{s}_1, \ldots, \mathbf{s}_k \in \mathbb{R}^2$*, the vector*

$$\begin{bmatrix} a(\mathbf{s}_1) & \cdots & a(\mathbf{s}_k) \end{bmatrix}$$

has a multivariate Gaussian distribution.

A Gaussian random field over $\mathbb{R}^2$ is described by a mean function, and by a covariance function $\mathcal{C}(\mathbf{s}, \mathbf{s}')$.

DEFINITION 5.2. *A* log-Gaussian random field $a(\mathbf{s})$ *over* $\mathbb{R}^2$ *is a random field whose natural logarithm* $\ln(a(\mathbf{s}))$ *is a Gaussian random field.*

Specifically, for any integer $k > 0$*, and any* k *fixed points* $\mathbf{s}_1, \ldots, \mathbf{s}_k \in \mathbb{R}^2$*, the vector*

$$\begin{bmatrix} \ln(a(\mathbf{s}_1)) & \cdots & \ln(a(\mathbf{s}_k)) \end{bmatrix}$$

has a multivariate Gaussian distribution.

Log-Gaussian fields are described by the mean function and the covariance function of the underlying Gaussian random field.

5.2. Original problem. As in [6, Section 5], we define the function f in terms of the numerical solution of a partial differential equation.

The PDE is

$$-\nabla_{\mathbf{s}} \cdot (a(\mathbf{s}) \nabla_{\mathbf{s}} u(\mathbf{s}, a(\mathbf{s}))) = 1, \qquad \mathbf{s} \in S \equiv [0,1] \times [0,1], \tag{5.1}$$

with boundary conditions $u = 0$. The coefficient $a(\mathbf{s})$ is a log-Gaussian random field with mean zero and covariance function $\mathcal{C}(\mathbf{s}, \mathbf{s}')$. The desired function f will be an approximation of $\int_S u(\mathbf{s}, a(\mathbf{s}))\, d\mathbf{s}$.

This log-Gaussian random field can be represented with a Karhunen-Loéve expansion [25, (5.5)],

$$\ln(a(\mathbf{s})) = \sum_{i=1}^{\infty} \sqrt{\mu_i}\, \phi_i(\mathbf{s})\, x_i,$$

where μ_i are the eigenvalues of the covariance $\mathcal{C}(\mathbf{s}, \mathbf{s}')$, $\phi_i(\mathbf{s})$ are the associated orthonormal eigenfunctions, and the scalars x_i are independent $\mathcal{N}(0,1)$ random variables.

Numerical solution. We solve (5.1) with a finite element discretization from MATLAB's PDE Toolbox[4], and approximate $\ln(a(\mathbf{s}))$ at the nodes $\mathbf{n}_i$, $1 \le i \le N$. The eigenvalues and eigenfunctions of the covariance function $\mathcal{C}(\mathbf{s}, \mathbf{s}')$ are approximated by the eigenvalues $\hat{\mu}_i$ and eigenvectors $\hat{\phi}_i$ of the $N \times N$ covariance matrix C with elements

$$C_{ij} = \mathcal{C}(\mathbf{n}_i, \mathbf{n}_j), \qquad 1 \le i \le N, 1 \le j \le N.$$

Small eigenvalues $|\hat{\mu}_i| < 10^{-12}$ are truncated. With $m \le N$ being the number of sufficiently large eigenvalues remaining, we approximate $\ln(a(\mathbf{s}))$ by

$$\hat{a}(\mathbf{x}) \equiv \sum_{i=1}^{m} \sqrt{\hat{\mu}_i}\, \hat{\phi}_i x_i, \qquad \text{where} \quad \mathbf{x} \equiv \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix},$$

and x_i are independent $\mathcal{N}(0,1)$ random variables.

[4] `http://www.mathworks.com/products/pde/`

Well-posedness. If $a(\mathbf{s}) \geq \alpha_{min} > 0$ for all $\mathbf{s} \in S$, then (5.1) is well-posed [27, pages 13, 28]. If also $0 < \alpha_{min} \leq a(\mathbf{s}) \leq \alpha_{max}$ for all $\mathbf{s} \in S$, then the error in the finite element solution can be bounded [11, page 128].

Since the Karhunen-Loéve expansion of $\ln a(\mathbf{s})$ depends on Gaussian random variables x_i, it is not possible to bound $a(\mathbf{s})$ from above and below. However, for a given fixed vector $\mathbf{x}$, one can determine, a posteriori, bounds so that $\alpha_{min} \leq \exp(\hat{a}(\mathbf{x})) \leq \alpha_{max}$. The largest such α_{max} and the smallest α_{min} are reported in Section 6.2.

Construction of f. The elements of the N-vector $\mathbf{u}$ are the solutions of the discretized PDE at the nodes $\mathbf{n}_i$, and we define f via

$$f(\mathbf{x}) \equiv \mathbb{1}^T M \, \mathbf{u},$$

where M is the $N \times N$ mass matrix of the finite element discretization, and $\mathbb{1}$ is the N-vector of all ones.

Since $\hat{a}(\mathbf{x})$ involves m random $\mathcal{N}(0,1)$ variables, $f(\mathbf{x})$ depends on m parameters. However the dimension of the active subspace appears to be much lower. It has been observed [6, Section 5.2] and [8, Section 5.2] that functions closely related to f are sensitive to change primarily along only a single direction in $\mathbb{R}^m$. Hence, for all practical purposes, f has an approximate active subspace of dimension $k = 1$.

5.3. Generalization. We generalize (5.1) to obtain functions f that are sensitive to change along several directions and give rise to approximate active subspaces of higher dimension.

To that end, consider the family of PDEs

$$-\nabla_{\mathbf{s}} \cdot (a(\mathbf{s}, w) \, \nabla_{\mathbf{s}} u(\mathbf{s}, a(\mathbf{s}, w)) = 1, \quad \mathbf{s} \in S = [0,1] \times [0,1], \quad 1 \leq w \leq W, \quad (5.2)$$

with boundary conditions $u = 0$ for each PDE. The wth coefficient $a(\mathbf{s}, w)$ is a log-Gaussian random field with mean zero and covariance function $\mathcal{C}_w(\mathbf{s}, \mathbf{s}')$. The desired function f will be an approximation of $\sum_{w=1}^{W} \int_S u\,(\mathbf{s}, a(\mathbf{s}, w))\, d\mathbf{s}$.

Numerical solution. As in Section 5.2, we compute a finite element discretization of (5.1) at N nodes n_i, and express each $\ln a(\mathbf{s}, w)$ in terms of a Karhunen-Loéve expansion, which gives a $N \times N$ covariance matrix with $m_w \leq N$ sufficiently large eigenvalues, $1 \leq w \leq W$.

The resulting approximation $\hat{a}_w(\mathbf{x}_w)$ of $\ln a(\mathbf{s}, w)$ depends on a vector $\mathbf{x}_w$, whose m_w elements are independent random $\mathcal{N}(0,1)$ variables, $1 \leq w \leq W$.

Construction of f. The elements of the N-vector $\mathbf{u}_w$ are the solutions of the wth discretized PDE at the nodes $\mathbf{n}_i$, and define f via

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_W) \equiv \mathbb{1}^T M \sum_{w=1}^{W} \mathbf{u}_w,$$

where M is the $N \times N$ mass matrix from Section 5.2.

The function f depends on a total of $m = \sum_{w=1}^{W} m_w$ parameters. If the W covariance functions $\mathcal{C}_w(\mathbf{s}, \mathbf{s}')$ are sufficiently different, then f should vary along W directions and have an approximate active subspace of dimension $k = W$. The identification of this subspace in Algorithm 1 requires computing gradients $\nabla f(\mathbf{x}_1, \ldots, \mathbf{x}_W)$. This is done by extending the method in [6, Section 5] from $W = 1$ to $W > 1$.

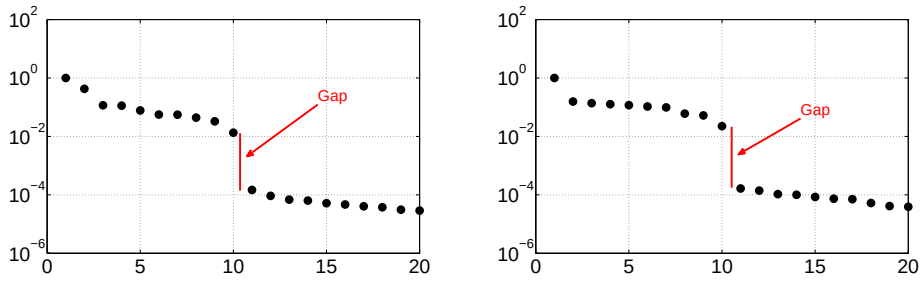

FIG. 6.1. *Twenty largest eigenvalue ratios* $\widehat{\lambda}_j/\widehat{\lambda}_1$ *of the* 3495×3495 *matrix* $\widehat{E}$ *versus index* j*, computed by Algorithm 1 with* n_1 *samples. Left plot:* $n_1 = 100$*. Right plot:* $n_1 = 1000$*.*

6. Numerical experiments. We evaluate the quality of the approximate response surfaces on the problem (5.2) with $W = 10$, by means of the error measures in Section 4.2

The covariance functions are based on two families, exponentials and rational quadratics see [22, page 86] and [1, Sections 4.2.3 and 4.2.4]),

$$\mathcal{C}(\mathbf{s}, \mathbf{s}') = \exp(-\|\mathbf{s} - \mathbf{s}'\|_2^{\alpha}) \quad \text{and} \quad \mathcal{C}(\mathbf{s}, \mathbf{s}') = \left(1 + \frac{\|\mathbf{s} - \mathbf{s}'\|_2^2}{2\,\alpha}\right)^{\alpha}$$

for $\alpha \in \{2/5, 4/5, 6/5, 8/5, 10/5\}$.

We solve the system (5.2) of 10 PDEs with MATLAB's PDE Toolbox, on a finite element mesh with $N = 712$ nodes. Due to the number of sufficiently large eigenvalues in the covariance matrices, $f(\mathbf{x}_1, \ldots, \mathbf{x}_{10})$ depends on a total of $m = \sum_{w=1}^{10} m_w = 3495$ parameters.

In the following we discuss: Identifying an approximate active subspace with Algorithm 1 (Section 6.1); computing training pairs with Algorithm 2 and constructing an approximate response surface (Section 6.2); and assessing the error between f and the approximate response surface (Section 6.3).

6.1. Identifying an approximate active subspace. We confirm that the function $f(\mathbf{x}_1, \ldots, \mathbf{x}_{10})$ is indeed sensitive to change along 10 directions in $\mathbb{R}^{3495}$. To this end, we compute the gaps between the dominant eigenvalues of the 3495×3495 matrix $\widehat{E}$ in (2.5).

Algorithm 1 represents $\widehat{E}$ in factored form as a Gram product $\widehat{E} = \frac{1}{n_1} GG^T$, where G is a $3495 \times n_1$ matrix, and computes the singular values of σ_j of G. The properly normalized squared singular values of G are eigenvalues of $\widehat{E}$, that is $\widehat{\lambda}_j = \sigma_j^2/n_1$.

Figure 6.1 shows the 20 largest eigenvalue ratios $\widehat{\lambda}_j/\widehat{\lambda}_1$ of $\widehat{E}$, computed by Algorithm 1 with two sampling amounts: $n_1 = 100$ and $n_1 = 1000$. Both amounts clearly produce a large relative gap $(\widehat{\lambda}_{10} - \widehat{\lambda}_{11})/\widehat{\lambda}_1 \approx 100$. Thus $f(\mathbf{x}_1, \ldots, \mathbf{x}_{10})$ has an approximate active subspace of dimension $k = 10$ in its parameter space $\mathbb{R}^{3495}$.

6.2. Computation of training pairs, and construction of approximate response surface. We use MATLAB's Sparse Grid Interpolation Toolbox [18, 19], which implements piecewise multilinear interpolation, to select training points and construct an approximate response surface.

Gaussian processes [6, Section 5] would have been an alternative option. They have the advantage of added "confidence intervals" around the constructed surfaces.

However, we decided in favour of sparse grid interpolation due to the availability of robust software, appropriateness for high-dimensional problems, and simplicity.

Though the goal is a response surface h over a subspace of dimension $k = 10$, we also construct surfaces over spaces of dimensions $1 \le k \le 14$, to get a sense for how the accuracy of h changes with increasing k. For efficiency, we replace the full space $\mathbb{R}^k$ by a k-dimensional hypercube $[-3,3]^k = [-3,3]\times\cdots\times[-3,3]$, which covers 3 standard deviations of $\mathcal{N}(0,1)$ random variables. The relative tolerance of the toolbox is set to 10^{-1}.

The toolbox constructs a sparse grid over $[-3,3]^k$. The resulting sparse grid points are the training points $\widehat{\mathbf{y}}_i$ and inputs for Algorithm 2. Table 6.1 displays the number of training points for subspaces of dimension $1 \le k \le 14$.

We make two simplifications in Algorithm 2.

1. Set $n_2 = 1$.
   This can be done, see also [6, Section 4.2], since the sampling amount n_2 has only a weak effect on the mean square error bound of h in Theorem 2.8.
2. Set $\widehat{\mathbf{z}}_1 = 0$.
   With only a single sample, it is not necessary to draw from the conditional distribution.

Regarding the boundedness discussed at the end of Section 5.2, the largest observed α_{max} and the smallest observed α_{min} values are 933 and 0.001, respectively.

6.3. Error of the approximate response surface. We evaluate the error of the approximate response surface h with the two measures discussed in Section 4.

Mean squared error. Theorem 2.8 bounds the averages of the absolute errors $(f(\mathbf{x}) - h(\hat{V}_1^T\mathbf{x}))^2$ at all $\mathbf{x} \in \mathbb{R}^{3495}$. Our numerical approximation

$$\frac{1}{10^5}\sum_{i=1}^{10^5}\left(f(\mathbf{x}_i) - h(\widehat{V}_1^T\mathbf{x}_i)\right)^2 \tag{6.1}$$

is based on 10^5 points $\mathbf{x}_i \in \mathbb{R}^{3495}$ whose elements are independent random $\mathcal{N}(0,1)$ variables.

Figure 6.2 shows the error (6.1) for approximate response surfaces h over active subspaces of dimension $1 \le k \le 14$. The error (6.1) decreases with increasing dimension, but starts to stagnate at $k = 10$, confirming that the approximate active subspace has dimension 10. Furthermore, (6.1) does not change substantially when the sampling amount of Algorithm 1 increases from $n_1 = 100$ to $n_1 = 1000$.

Pointwise relative error. First we examine the errors at testing points inside the approximate active subspace, and then at points outside.

k	1	2	3	4	5	6	7
Training points	5	29	177	1105	6993	15121	30241
k	8	9	10	11	12	13	14
Training points	56737	100897	171425	280017	442001	677041	1009905

TABLE 6.1

Number of sparse grid points in $[-3,3]^k$ versus k, for $1 \le k \le 14$. The points serve as the training points for Algorithm 2.

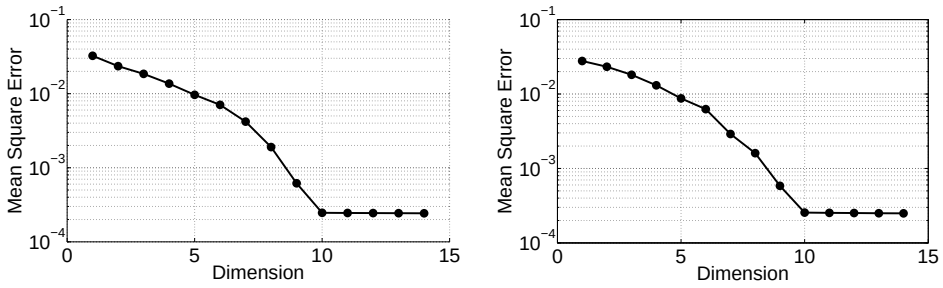

FIG. 6.2. *Approximate MSE (6.1) versus dimension k, for $1 \le k \le 14$. Approximate active subspaces for h are computed by Algorithm 1 with n_1 samples. Left plot: $n_1 = 100$. Right plot: $n_1 = 1000$.*

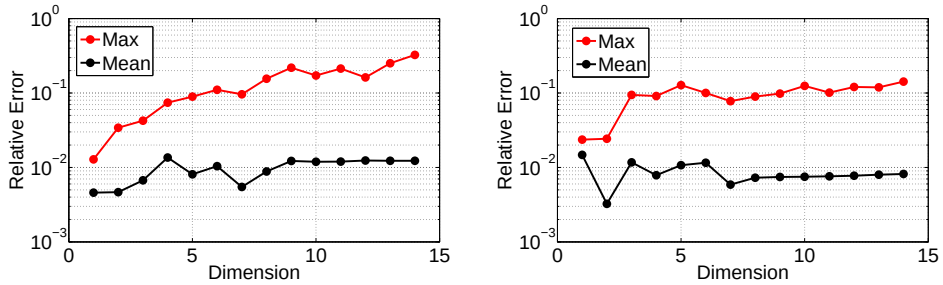

FIG. 6.3. *Response surface accuracy inside approximate active subspaces: Maximum and mean relative errors (6.2) versus dimension k for $1 \le k \le 14$. Approximate active subspaces are computed by Algorithm 1 with n_1 samples. Left plot: $n_1 = 100$. Right plot: $n_1 = 1000$.*

Testing points in the approximate active subspace. We determine the relative errors (4.1) at testing points in $\mathbf{x}_i \in \mathsf{range}(\widehat{V}_1)$,

$$\left| \frac{h(\widehat{V}_1^T \mathbf{x}_i) - f(\mathbf{x}_i)}{f(\mathbf{x}_i)} \right| \quad \text{for} \quad \mathbf{x}_i \in \mathsf{range}(\widehat{V}_1) \subset [-3,3]^k, \quad 1 \le i \le 1000k. \tag{6.2}$$

A testing point is computed as $\mathbf{x}_i = \widehat{V}_1 \mathbf{z}$, where the elements of $\mathbf{z} \in \mathbb{R}^k$ are independent random $\mathcal{U}(-3,3)$ variables.

Figure 6.3 displays the maximum and mean relative errors (6.2). For sampling amounts $n_1 = 100$ in Algorithm 1, the mean is about 10^{-2}. With larger amounts $n_1 = 1000$, there is a slight decrease in the maximum and mean relative errors associated with the larger dimensions $8 \le k \le 14$. We were disappointed, though, to see relative errors exceeding the relative interpolation tolerance of 10^{-1}.

To understand the distribution of errors for the dimension of interest, $k = 10$, we plot all 10000 relative errors in Figure 6.4. For both sampling amounts, most errors are below 10^{-1}. Furthermore, for $n_1 = 100$, about 60% of the relative errors are below 10^{-2}, and for $n_1 = 1000$ this increases to about 80%.

Testing points outside the approximate active subspace.. We determine the relative errors (4.1) at testing points outside $\mathbf{x}_i \in \mathsf{range}(\widehat{V}_1)$,

$$\left| \frac{h(\widehat{V}_1^T \mathbf{x}_i) - f(\mathbf{x}_i)}{f(\mathbf{x}_i)} \right| \quad \text{for} \quad \mathbf{x}_i \notin \mathsf{range}(\widehat{V}_1), \quad 1 \le i \le 1000k. \tag{6.3}$$

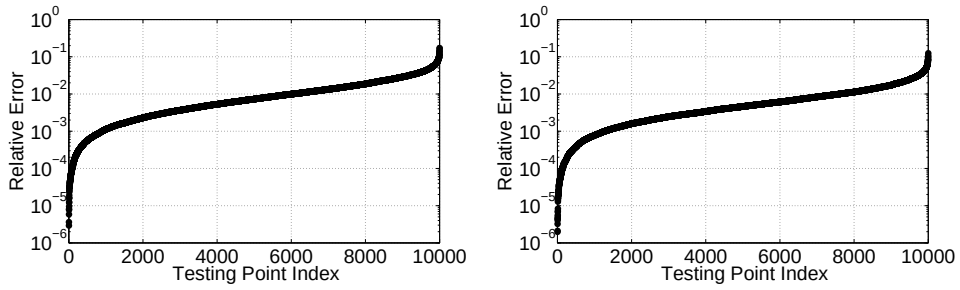

FIG. 6.4. *Response surface accuracy inside the approximate active subspace of dimension* $k = 10$*: Sorted relative errors (6.2) versus testing point index* i*. Approximate active subspaces are computed by Algorithm 1 with* n_1 *samples. Left plot:* $n_1 = 100$*. Right plot:* $n_1 = 1000$*.*

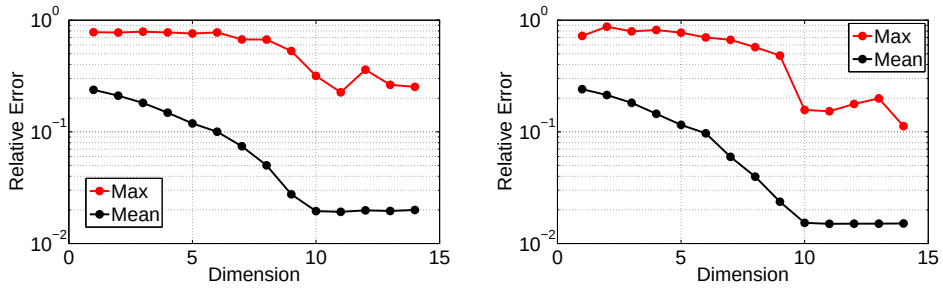

FIG. 6.5. *Response surface accuracy outside approximate active subspaces: Maximum and mean relative errors (6.3) versus dimension* k *for* $1 \le k \le 14$*. Approximate active subspace are computed by Algorithm 1 with* n_1 *samples. Left plot:* $n_1 = 100$*. Right plot:* $n_1 = 1000$*.*

A testing point is computed as $\mathbf{x}_i = \widehat{V}_1 \mathbf{z}_1 + (I - \widehat{V}_1 \widehat{V}_1^T)\, \mathbf{z}_2$, where the elements of $\mathbf{z}_1 \in \mathbb{R}^k$ are independent random $\mathcal{U}(-3, 3)$ variables, and the elements of $\mathbf{z}_2 \in \mathbb{R}^{m-k}$ are independent random $\mathcal{N}(0, 1)$ variables.

Figure 6.5 displays the maximum and mean relative errors (6.3). For sampling amounts $n_1 = 100$ in Algorithm 1, the mean error decreases with increasing dimension, until it stagnates at about 10^{-2} for dimension $k = 10$. With larger amounts $n_1 = 1000$, there is a slight decrease in the mean. The maximum relative errors exceed 10^{-1} for all dimensions $1 \le k \le 14$ and both sampling amounts n_1.

To understand the distribution of relative errors for the dimension of interest, $k = 10$, we plot all relative errors. The maximum does exceed 10^{-1} for both sampling amounts n_1, but most errors are below 10^{-1}. Overall, larger smaller sampling amounts result in slightly smaller relative errors.

7. Conclusions. The numerical experiments, as illustrated in Figures 6.2 – 6.6, suggest that approximate response surfaces over subspaces of dimension $k = 10$ can be approximated, without too much effort, to an absolute accuracy of at least 10^{-3} and a relative accuracy of at least 10^{-1}.

Note, though, that the experiments in Section 6 were computationally feasible because the active subspace has the small dimension $k = 10$. For problems with high-dimensional active subspaces, however, say $k = 10^5$ or $k = 10^6$, the SVD-based subspace computation in Algorithm 1 is too expensive. We believe there is a lot more potential for randomized algorithms [14, 20], in particular when it comes to replacing the SVD in Algorithm 1 by a randomized iterative method for approximation of

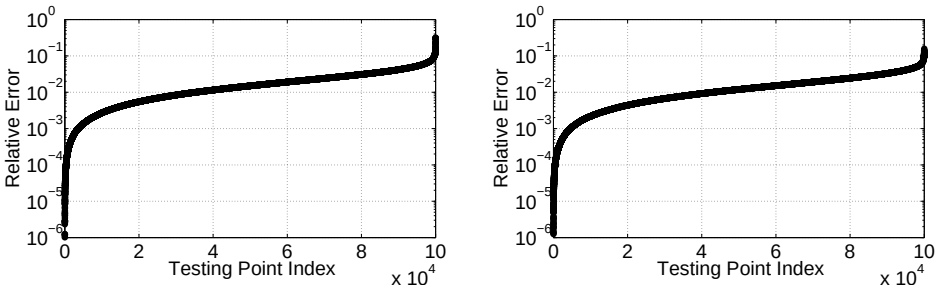

FIG. 6.6. *Response surface accuracy outside the approximate active subspaces of dimension $k = 10$: Sorted relative errors (6.3) versus testing point index i. Approximate active subspace are computed by Algorithm 1 with n_1 samples. Left plot: $n_1 = 100$. Right plot: $n_1 = 1000$.*

dominant subspaces.

8. Acknowledgements. We thank Tim Kelley and Paul Constantine for helpful discussions, and Paul Constantine also for the MATLAB codes from [6, Section 5].

Appendix A. Auxiliary results and proof of Theorem 3.1. We present a number of auxiliary results in Section A.1 that are required for the proof of Theorem 3.1 in Section A.2.

A.1. Auxiliary results. We start with a matrix concentration inequality (Theorem A.2), which provides the basis for a probabilistic bound (Theorem A.3) on the relative error $\epsilon = \|\widehat{E} - E\|_2/\|E\|_2$ in terms of the number of samples n_1 used by Algorithm 1. This can be rearranged (Corollary A.4) into a lower bound on the number of samples n_1 in terms of ϵ. At last, a "structural" linear algebra result makes the transition from ϵ to the angle between the approximate and ideal active subspaces.

The concentration inequality depends on the following version of "rank".

DEFINITION A.1 (Definition 7.1.1 in [31]). *If the $m \times m$ matrix P is real symmetric positive semi-definite, then its intrinsic dimension is*

$$\mathsf{intdim}\,(P) \equiv \mathsf{trace}\,(P)\,/\,\|P\|_2\,,$$

where $1 \leq \mathsf{intdim}\,(P) \leq \mathsf{rank}\,(P) \leq m$.

The following matrix concentration inequality bounds a sum of independent random matrices. In this specialization of the matrix Bernstein inequality [31, Theorem 7.3.1] to symmetric matrices, the "random matrices" are matrix-valued random variables that are bounded with zero mean, and whose "variance" is bounded, in the sense of the Löwner partial order[5], by a symmetric positive semi-definite matrix P. Below is a specialization

THEOREM A.2 (Theorem 7.3.1 in [31]). *Suppose that*

1. *X_j are n_1 independent real symmetric random matrices,*
2. *$\mathbf{E}[X_j] = 0$ and $\|X_j\|_2 \leq p_1$ for $1 \leq j \leq n_1$,*
3. *P is a real symmetric positive semi-definite matrix so that $P - \sum_{j=1}^{n_1} \mathbf{E}[X_j^2]$ is positive semi-definite.*

[5]If P_1 and P_2 are real symmetric matrices, then $P_1 \preceq P_2$ means that $P_2 - P_1$ is positive semi-definite [16, Definition 7.7.1].

If $\epsilon \geq \|P\|_2^{1/2} + p_1/3$, *then*

$$\textbf{Prob}\left[\left\|\sum_{j=1}^{n_1} X_j\right\|_2 \geq \epsilon\right] \leq 4\,\mathsf{intdim}\,(P)\,\exp\left(\frac{-\epsilon^2/2}{\|P\|_2 + p_1\epsilon/3}\right).$$

Since $\mathbf{E}\left[\sum_{j=1}^{n_1} X_j\right] = 0$, Theorem A.2 bounds the deviation of the sum from its mean. The bound for the probability does not depend on the dimension of the random variables. Theorem A.2 is now applied to the matrices in Sections 2.1 and 2.2.

THEOREM A.3. *Let*

$$E \equiv \int_{\mathbb{R}^m} \nabla f(\mathbf{x})\,(\nabla f(\mathbf{x}))^T\,\rho(\mathbf{x})\,d\mathbf{x} \quad \textit{and} \quad \widehat{E} \equiv \frac{1}{n_1}\sum_{j=1}^{n_1} \nabla f(\mathbf{x}_j)(\nabla f(\mathbf{x}_j))^T,$$

where $\mathbf{x}_j$ *are sampled independently according to* $\rho(\mathbf{x})$*, as in Algorithm 1. If* $0 < \delta < 1$*, then with probability at least* $1-\delta$*,*

$$\frac{\left\|\widehat{E} - E\right\|_2}{\|E\|_2} \leq \widehat{\gamma} + \sqrt{\widehat{\gamma}(\widehat{\gamma}+6)} \quad \textit{where} \quad \widehat{\gamma} \equiv \frac{L^2}{3\,n_1\,\|E\|_2}\,\ln\left(\frac{4}{\delta}\,\mathsf{intdim}\,(E)\right).$$

Proof. The proof is similar to the one of [15, Theorem 7.8].

Write $\widehat{E} - E = \sum_{j=1}^{n_1} X_j$, where

$$X_j \equiv \frac{1}{n_1}\nabla f(\mathbf{x}_j)(\nabla f(\mathbf{x}_j))^T - \frac{1}{n_1}E, \qquad 1 \leq j \leq n_1.$$

In order to apply Theorem A.2 to the above sum, we need to verify the assumptions.

Zero mean. From the definition of E and the fact that ρ is a probability density function follows

$$\begin{aligned}\mathbf{E}[X_j] &= \frac{1}{n_1}\int (\nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T - E)\rho(\mathbf{x})\,d\mathbf{x} \\ &= \frac{1}{n_1}\int \nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T\rho(\mathbf{x})\,d\mathbf{x} - \frac{E}{n_1}\int \rho(\mathbf{x})\,d\mathbf{x} = 0.\end{aligned}$$

Boundedness. The Lipschitz continuity of f implies

$$\|X_j\|_2 \leq \frac{1}{n_1}\,\max\left\{\left\|\nabla f(\mathbf{x}_j)(\nabla f(\mathbf{x}_j))^T\right\|_2,\ \|E\|_2\right\} \leq \frac{1}{n_1}\max\{L^2, \|E\|_2\}.$$

Assumptions 1.1 ensure that the second term in the maximum can be bounded by

$$\begin{aligned}\|E\|_2 &= \left\|\int \nabla f(\mathbf{x}))(\nabla f(\mathbf{x}))^T\rho(\mathbf{x})\,d\mathbf{x}\right\|_2 \\ &\leq \int \left\|\nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T\right\|\rho(\mathbf{x})\,d\mathbf{x} \leq L^2.\end{aligned}$$

Thus

$$\|X_j\|_2 \leq p_1 \equiv \frac{L^2}{n_1}, \qquad 1 \leq j \leq n_1. \tag{A.1}$$

"Variance". Multiply out the integrand and apply the definition of E,

$$\begin{aligned}\mathbf{E}[X_j^2] &= \frac{1}{n_1^2}\int \left(\nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T - E\right)^2 \rho(\mathbf{x})\,d\mathbf{x} \\ &= \frac{1}{n_1^2}\left[\int (\nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T)^2 \rho(\mathbf{x})\,d\mathbf{x} - E^2 - E^2 + E^2\right] \\ &= \frac{1}{n_1^2}\left[\int (\nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T)^2 \rho(\mathbf{x})\,d\mathbf{x} - E^2\right].\end{aligned}$$

The positive semi-definiteness of E^2 implies $0 \preceq E^2$, hence

$$\mathbf{E}[X_j^2] \preceq \frac{1}{n_1^2}\int (\nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T)^2 \rho(\mathbf{x})\,d\mathbf{x}. \tag{A.2}$$

Bounding the "variance". Since $\nabla f(\mathbf{x})$ is a column vector, the squared outer product contains an inner product,

$$\int (\nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T)^2 \rho(\mathbf{x})\,d\mathbf{x} = \int \|\nabla f(\mathbf{x})\|_2^2\, \nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T \rho(\mathbf{x})\,d\mathbf{x}.$$

From (A.2) follows that this matrix is positive semi-definite. This, together with Assumptions 1.1 and the Lipschitz continuity of f implies

$$\frac{1}{n_1^2}\int \|\nabla f(\mathbf{x})\|_2^2\, \nabla f(\mathbf{x})(\nabla f(\mathbf{x}))^T \rho(\mathbf{x})\,d\mathbf{x} \preceq \frac{L^2}{n_1^2}\,E.$$

Combine this with (A.2) to conclude $\mathbf{E}\left[X_j^2\right] \preceq \frac{L^2}{n_1^2}\,E \preceq \frac{L^2}{n_1}\,E$, and set $P \equiv \frac{L^2}{n_1}\,E$.

The linearity of trace and norm implies

$$\|P\|_2 = \frac{L^2}{n_1}\,\|E\|_2, \qquad \mathsf{intdim}\,(P) = \frac{\mathsf{trace}\,(E)}{\|E\|_2}. \tag{A.3}$$

Application of Theorem A.2. Substituting (A.1) and (A.3) into the bound of Theorem A.2 gives

$$\mathbf{Prob}\left[\left\|\widehat{E} - E\right\|_2 \geq \hat{\epsilon}\right] \leq 4\,\mathsf{intdim}\,(E)\,\exp\left(\frac{-\hat{\epsilon}^2/2}{L^2/n_1\,\|E\|_2 + L^2\,\hat{\epsilon}/(3\,n_1)}\right).$$

Setting the above right hand side equal to δ and solving for $\hat{\epsilon}$ gives

$$\hat{\epsilon} = \gamma + \sqrt{\gamma(\gamma + 6\,\|E\|_2)}$$

where

$$\gamma \equiv \frac{L^2}{3\,n_1}\,\ln\left(\frac{4}{\delta}\,\mathsf{intdim}\,(E)\right).$$

Check lower bound for $\hat{\epsilon}$. We need to verify that the quantities above satisfy $\hat{\epsilon} \geq p_1/3 + \|P\|_2^{1/2}$. From $0 < \delta < 1$ and $\mathsf{intdim}\,(E) \geq 1$ follows $e < \frac{4}{\delta} \leq \frac{4}{\delta}\,\mathsf{intdim}\,(E)$, hence $\ln(\frac{4}{\delta}\,\mathsf{intdim}\,(E)) \geq 1$. This implies

$$\frac{p_1}{3} = \frac{L^2}{3\,n_1} \leq \frac{L^2}{3\,n_1}\,\ln\left(\frac{4}{\delta}\,\mathsf{intdim}\,(E)\right) = \gamma$$

and

$$\|P\|_2^{1/2} = \sqrt{\frac{\|E\|_2}{n_1}}\,L \leq \sqrt{6\gamma\,\|E\|_2} \leq \sqrt{\gamma(\gamma + 6\,\|E\|_2)}.$$

Adding the two inequalities gives the lower bound for $\hat{\epsilon}$.

Relative Error. Dividing both sides of $\left\|\widehat{E}-E\right\|_2 \leq \hat{\epsilon}$ by $\|E\|_2$ gives

$$\frac{\left\|\widehat{E}-E\right\|_2}{\|E\|_2} \leq \frac{\hat{\epsilon}}{\|E\|_2} = \widehat{\gamma} + \sqrt{\widehat{\gamma}(\widehat{\gamma}+6)} \qquad \text{where} \quad \widehat{\gamma} = \gamma / \|E\|_2 .$$

□

The previous result implies a lower bound on the number of samples n_1 required for a user-specified relative error.

COROLLARY A.4. *If, in addition to the conditions of Theorem A.3, also* $0 < \epsilon < 1$ *and*

$$n_1 \geq \frac{8}{3}\,\frac{L^2}{\epsilon^2\,\|E\|_2}\ln\left(\frac{4}{\delta}\,\mathsf{intdim}\,(E)\right),$$

then with probability at least $1-\delta$

$$\frac{\|\widehat{E}-E\|_2}{\|E\|_2} \leq \epsilon.$$

Proof. We want to determine n_1 such that $\widehat{\gamma}$ in Theorem A.3 satisfies

$$\widehat{\gamma} + \sqrt{\widehat{\gamma}(\widehat{\gamma}+6)} \leq \epsilon. \tag{A.4}$$

Set $t \equiv \frac{L^2}{\|E\|_2}\ln\left(\frac{4}{\delta}\,\mathsf{intdim}\,(E)\right)$, so that $\widehat{\gamma} = t/(3\,n_1)$ and $n_1 = \alpha\,t/\epsilon^2$ for some α. Then (A.4) is equivalent to determining α such that

$$\frac{\epsilon^2}{3\,\alpha} + \sqrt{\frac{\epsilon^2}{3\,\alpha}\left(\frac{\epsilon^2}{3\,\alpha}+6\right)} \leq \epsilon.$$

This inequality is satisfied by $\alpha \geq \frac{8}{3} \geq 2 + \frac{2}{3}\,\epsilon$. Now substitute this bound for α into the above expression for n_1. □

The final auxiliary result relates the absolute error $\|\widehat{E}-E\|_2$ to the distance between approximate and ideal active subspaces, provided $\|\widehat{E}-E\|_2$ is sufficiently small compared to the relevant eigenvalue gap. The bound below can viewed as a 2-norm version of [12, Corollary 8.1.11].

THEOREM A.5. *Let* E *and* $\widehat{E}$ *be* $m \times m$ *real symmetric matrices with respective eigenvalue decompositions (2.2) and (2.5) and partitioned as in (2.3) and (2.6). If* $\lambda_k - \lambda_{k+1} > 0$ *and*

$$\|\widehat{E}-E\|_2 \leq \frac{\lambda_k - \lambda_{k+1}}{4},$$

then

$$\left\|V_1V_1^T - \widehat{V}_1\widehat{V}_1^T\right\|_2 \leq 4\,\frac{\|\widehat{E}-E\|_2}{\lambda_k - \lambda_{k+1}}.$$

Proof. As in (2.2) partition

$$\begin{pmatrix} V_1 & V_2 \end{pmatrix}^T\, E\, \begin{pmatrix} V_1 & V_2 \end{pmatrix} = \begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix},$$

and set $gap \equiv \min_{i,j} |(\Lambda_1)_{ii} - (\Lambda_2)_{jj}| = \lambda_k - \lambda_{k+1}$. Also,

$$F = \begin{pmatrix} F_{11} & F_{12} \\ F_{12}^T & F_{22} \end{pmatrix} \equiv \begin{pmatrix} V_1 & V_2 \end{pmatrix}^T \, (\widehat{E} - E) \, \begin{pmatrix} V_1 & V_2 \end{pmatrix},$$

so that $\|F\|_2 = \|\widehat{E} - E\|_2$. When specialized to real symmetric matrices and the two-norm, [28, Theorems 2.7 and 4.11] or [29, page 232 and Theorem V.2.7] imply the following: If

$$\hat{\delta} \equiv gap - \|F_{11}\|_2 - \|F_{22}\|_2 > 0 \qquad \frac{\|F_{12}\|_2}{gap} < \tfrac{1}{2} \tag{A.5}$$

then

$$\|V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T\|_2 \leq 2 \, \|F_{12}\|_2 / \hat{\delta}. \tag{A.6}$$

Now let $\|F\|_2 < gap/4$. With $\delta \equiv gap - 2\|F\|_2$ conditions (A.5) hold, that is, $\hat{\delta} \geq \delta > \frac{1}{2} gap > 0$ and

$$\frac{\|F_{12}\|_2}{\hat{\delta}} \leq \frac{\|F\|_2}{\delta} \leq 2 \, \frac{\|F\|_2}{gap} < \tfrac{1}{2}.$$

The conclusion (A.6) holds with

$$\|V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T\|_2 \leq 2 \, \frac{\|F_{12}\|_2}{\hat{\delta}} \leq 4 \, \frac{\|F\|_2}{gap}.$$

□

At last we have all the ingredients to prove the desired result.

A.2. Proof of Theorem 3.1. From Corollary A.4 and the assumption $0 < \epsilon \leq \frac{\lambda_k - \lambda_{k+1}}{5 \, \lambda_1}$ follows with $\|E\|_2 = \lambda_1$: If

$$n_1 \geq \frac{8}{3} \, \frac{L^2}{\lambda_1 \, \epsilon^2} \ln \left(\frac{4}{\delta} \, \mathsf{intdim}\,(E) \right),$$

then with probability at least $1 - \delta$

$$\|\widehat{E} - E\|_2 \leq \|E\|_2 \, \epsilon \leq \frac{\lambda_k - \lambda_{k+1}}{4}.$$

Hence the assumptions of Theorem A.5 are satisfied, and with probability at least $1 - \delta$

$$\left\| V_1 V_1^T - \widehat{V}_1 \widehat{V}_1^T \right\|_2 \leq 4 \, \frac{\|\widehat{E} - E\|_2}{\lambda_k - \lambda_{k+1}} \leq \frac{4 \, \lambda_1 \, \epsilon}{\lambda_k - \lambda_{k+1}}.$$

REFERENCES

[1] P. Abrahamsen, *A review of Gaussian random fields and correlation functions*, Norwegian Computing Center, 2nd ed., 1997.

[2] Y. Bang, H. S. Abdel-Khalik, and J. M. Hite, *Hybrid reduced order modeling applied to nonlinear models*, Internat. J. Numer. Methods Engrg., 91 (2012), pp. 929–949.

[3] M. Bauerheim, A. Ndiaye, P. Constantine, G. Iaccarino, S. Moreau, and F. Nicoud, *Uncertainty quantification of thermo-acoustic instabilities in annular combustors*, tech. report, Center for Turbulence Research, Stanford University, 2014.

[4] H. Chen, Wang Q., R. Hu, and Constantine P. G., *Conditional sampling and experiment design for quantifying manufacturing error of transonic airfoil*, in 49th AIAA Aerospace Sciences Meeting, 2011.

[5] P. G. Constantine, A. Doostan, Q. Wang, and G. Iaccarino, *A surrogate accelerated Bayesian inverse analysis of the HyShot II flight data*, in Proceedings of the 52nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Material Conference, 2011.

[6] P. G. Constantine, E. Dow, and Q Wang, *Active subspace methods in theory and practice: Applications to Kriging surfaces*, SIAM J. Sci. Comput., 36 (2014), pp. A1500–A1524.

[7] P. G. Constantine, E. Dow, and Q. Wang, *Erratum: Active subspace methods in theory and practice: Applications to Kriging surfaces*, SIAM J. Sci. Comput., 36 (2014), pp. A3030–A3031.

[8] P. G. Constantine and D. Gleich, *Computing active subspaces with Monte Carlo*, 2015. arXiv:1408.0545v2.

[9] P. G. Constantine, Q. Wang, and G. Iaccarino, *A method for spatial sensitivity analysis*, tech. report, Center for Turbulence Research, Stanford University, 2012.

[10] P. G. Constantine, B. Zaharatos, and M. Campanelli, *Discovering an active subspace in a single-diode solar cell model*, 2014. arXiv:1406.7607.

[11] C. A. J. Fletcher, *Computational Galerkin Methods*, Springer Series in Computational Physics, Springer-Verlag, New York, 1984.

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, fourth ed., 2013.

[13] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, Oxford University Press, New York, third ed., 2001.

[14] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.

[15] J. T. Holodnak and I. C. F. Ipsen, *Randomized Approximation of the Gram Matrix: Exact Computation and Probabilistic Bounds*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 110–137.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, second ed., 2013.

[17] I. T. Jolliffe, *Principal component analysis*, Springer Series in Statistics, Springer-Verlag, New York, second ed., 2002.

[18] A. Klimke, *Sparse Grid Interpolation Toolbox – User's guide*, Tech. Report IANS report 2007/017, University of Stuttgart, 2007.

[19] A. Klimke and B. Wohlmuth, *Algorithm 847: spinterp: Piecewise multilinear hierarchical sparse grid interpolation in MATLAB*, ACM Transactions on Mathematical Software, 31 (2005).

[20] M. W. Mahoney, *Randomized algorithms for matrices and data*, Found. Trends Mach. Learning, 3 (2011), pp. 123–224.

[21] N. Namura, K. Shimoyama, and S. Obayashi, *Kriging surrogate model enhanced by coordinate transformation of design space based on eigenvalue decomposition*, in Evolutionary Multi-Criterion Optimization, Lecture Notes in Computer Science, Springer International Publishing, 2015.

[22] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2006.

[23] S. Ross, *Introduction to Probability Models*, Elsevier, Burlington, MA, 10th ed., 2010.

[24] T. M. Russi, *Uncertainty Quantification with experimental data and complex system models*, PhD thesis, University of California, Berkeley, 2010.

[25] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*, SIAM, Philadelphia, PA, 2014.

[26] I. M. Sobol' and S. Kucherenko, *Derivative based global sensitivity measures and their link with global sensitivity indices*, Math. Comput. Simulation, 79 (2009), pp. 3009–3017.

[27] P. Šolín, *Partial differential equations and the finite element method*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006.

[28] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–64.

[29] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego,

1990.
[30] M. Stoyanov and C. G. Webster, *A gradient-based sampling approach for dimension reduction of partial differential equations with stochastic coefficients*, Int. J. Uncertainty Quantification, (2014).
[31] J. A. Tropp, *An introduction to matrix concentration inequalities*, Found. Trends Mach. Learning, 8 (2015), pp. 1–230.